Hypothesis Testing - Frequentist

Compare two hypotheses to see which one better explains the data.
Or, alternatively, what is the best way to separate events into two classes, those originating from each of two hypotheses.
The two hypotheses are traditionally called:

$$H_0 : \text{the null hypothesis, and}$$
$$H_1 : \text{the alternative hypothesis.}$$

As usual, we know $P(\text{data}|H_0)$ and $P(\text{data}|H_1)$.

If W is the space of all possible data, we must find a
Critical Region $w \in W$ (in which we reject H_0) such that

$$P(\text{data} \in w \,|\, H_0) = \alpha \text{ (chosen to be small),}$$

and at the same time,

$$P(\text{data} \in (W - w) \,|\, H_1) = \beta \text{ is made as small as possible.}$$

Simple and Composite Hypotheses

Simple Hypotheses

When the hypotheses H_0 and H_1 are completely specified (with no free parameters), they are called simple hypotheses.

The theory of hypothesis testing for simple hypotheses is well developed, holds for large or small samples.

Composite Hypotheses

When a hypothesis contains one or more free parameters, it is a composite hypothesis, for which there is only an asymptotic theory.
Unfortunately, real problems usually involve composite hypotheses.

Fortunately, we can get exact answers for small samples using Monte Carlo.

Errors of first and second kinds

The level of significance α, (*size of test*) is defined as the probability of X falling in w (rejecting H_0) when H_0 is true: $P(X \in w \,|\, H_0) = \alpha$.

		H_0 TRUE	H_1 TRUE
$X \notin w$	ACCEPT H_0	Acceptance good Prob $= 1 - \alpha$	Contamination Error of the second kind Prob $= \beta$
$X \in w$ (critical region)	REJECT H_0	Loss Error of the first kind Prob $= \alpha$	Rejection good Prob $= 1 - \beta$

Errors of first and second kinds

The usefulness of a test depends on its ability to discriminate against the alternative hypothesis H_1. The measure of this usefulness is the power of the test, defined as the probability $1-\beta$ of X falling into the critical region if H_1 is true:

$$P(X \in w \mid H_1) = 1 - \beta\,. \tag{1}$$

In other words, β is the probability that X will fall in the acceptance region for H_0 if H_1 is true:

$$P(X \in W - w \mid H_1) = \beta\,.$$

In practice, the determination of a multidimensional critical region may be difficult, so one often chooses a single test statistic $t(X)$ instead. Then the critical region becomes w_t and we have:

$$P(t \in w_t \mid H_0) = \alpha$$

Example: Separation of two classes of events

Problem: Distinguish elastic proton scattering events

$$pp \to pp \quad \text{(the hypothesis under test, } H_0\text{)}$$

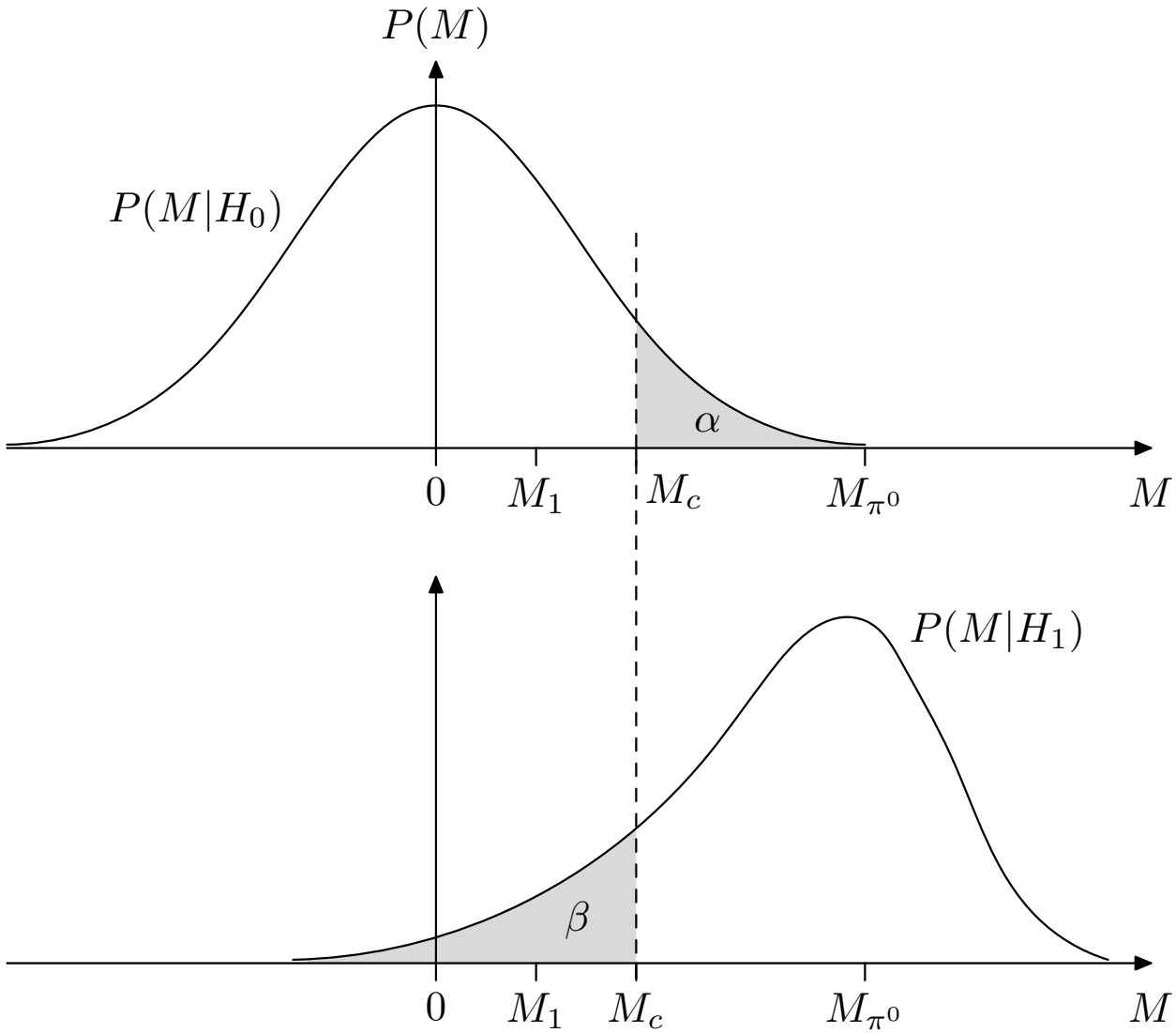
from inelastic scattering events

$$pp \to pp\pi^\circ \quad \text{(the alternative hypothesis, } H_1\text{)}\,,$$

in an experimental set-up which measures the proton trajectories, but where the π^0 cannot be detected directly. The obvious choice for the test statistic is the missing mass (the total rest energy of all unseen particles) for the event.

Knowing the expected distributions of missing mass for each of the two hypotheses, we can obtain the curves shown in the next slide which allow us to choose a critical region for missing mass M:

Example: Separation of two classes of events



Resolution functions for the missing mass M under the hypotheses H_0 and H_1, with critical region $M > M_c$.

Comparison of Tests: Power functions

If the two hypotheses under test can be expressed as two values of some parameter θ:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

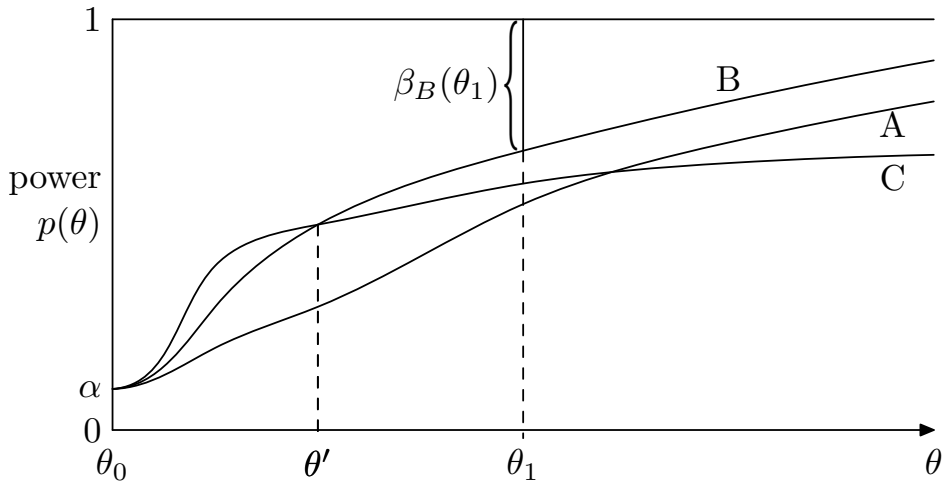
and the power of the test is $p(\theta_1) = P(X \in w|\theta_1) = 1 - \beta$.

It is of interest to consider $p(\theta)$, as a function of $\theta_1 = \theta$.

[Here H_1 is still considered a simple hypothesis, with θ fixed, but at different values.]

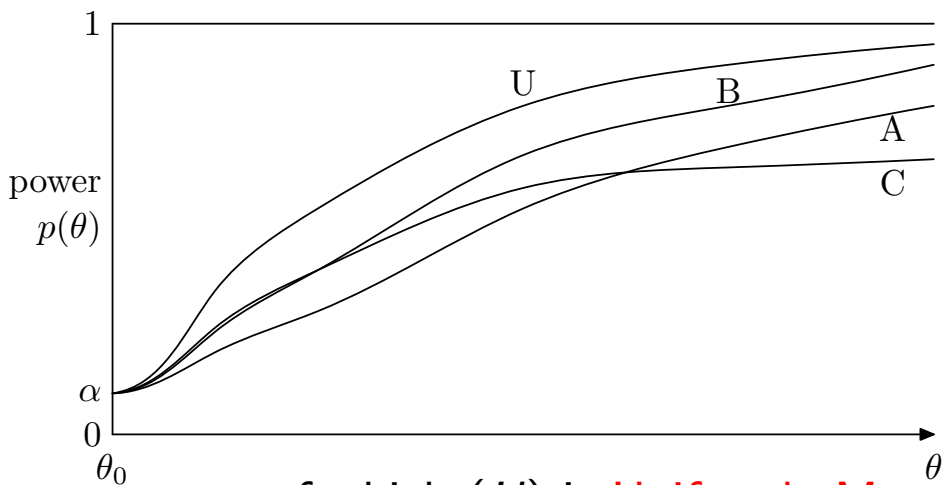
$p(\theta)$ is called a power function.

Comparison of Tests: Most powerful



Power functions of tests A, B, and C at significance level α.

Of these tests, B is the best for $\theta > \theta'$. For smaller values of θ, C is better.



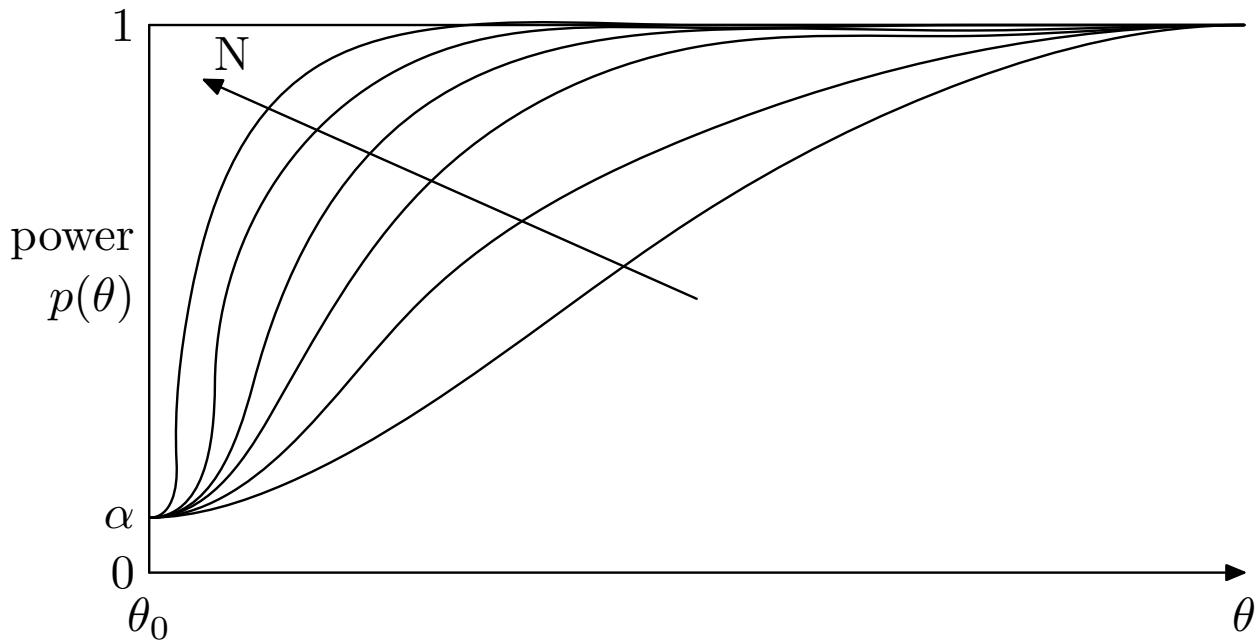
Power functions of four tests, one of which (U) is Uniformly Most Powerful.

Tests of Hypotheses: Consistency

A test is said to be consistent if the power tends to unity as the number of observations tends to infinity:

$$\lim_{N\to\infty} P(\mathbf{X} \in w_\alpha | H_1) = 1\,,$$

where $\mathbf{X}$ is the set of N observations, and w_α is the critical region, of size α, under H_0.

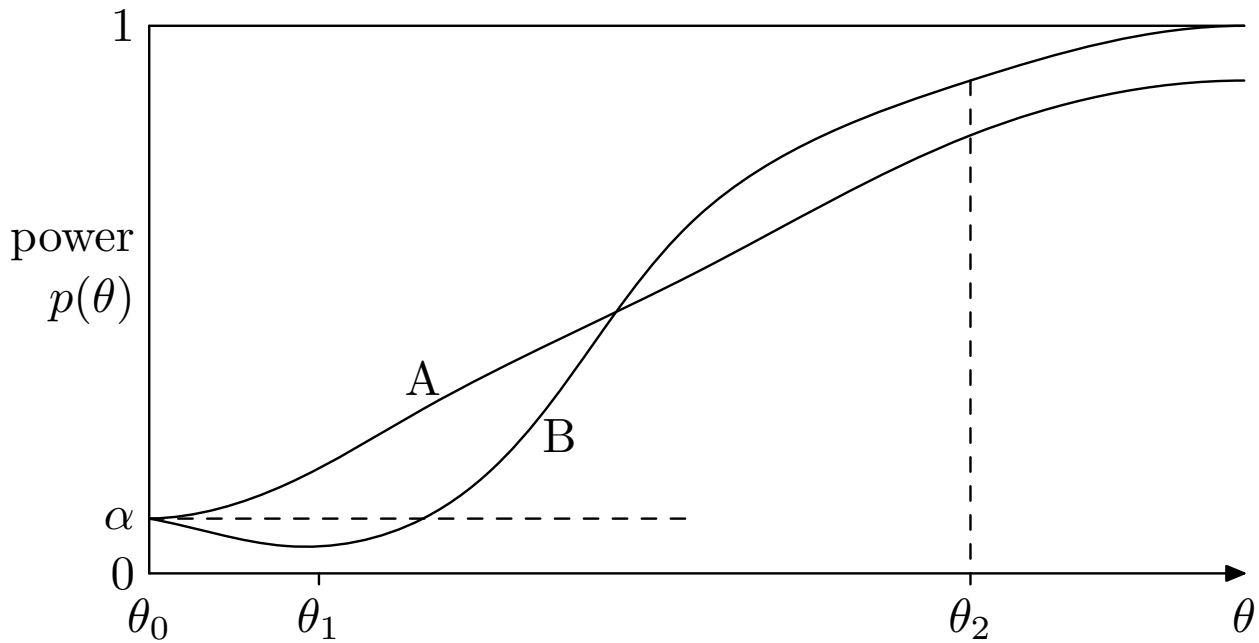


Power function for a consistent test as a function of N. As N increases, it tends to a step function.

Tests of Hypotheses: Bias

Consider the power curve which does not take on its minimum at $\theta = \theta_0$. In this case, the probability of accepting $H_0 : \theta = \theta_0$ is greater when $\theta = \theta_1$ than when $\theta = \theta_0$, or $1-\beta < \alpha$

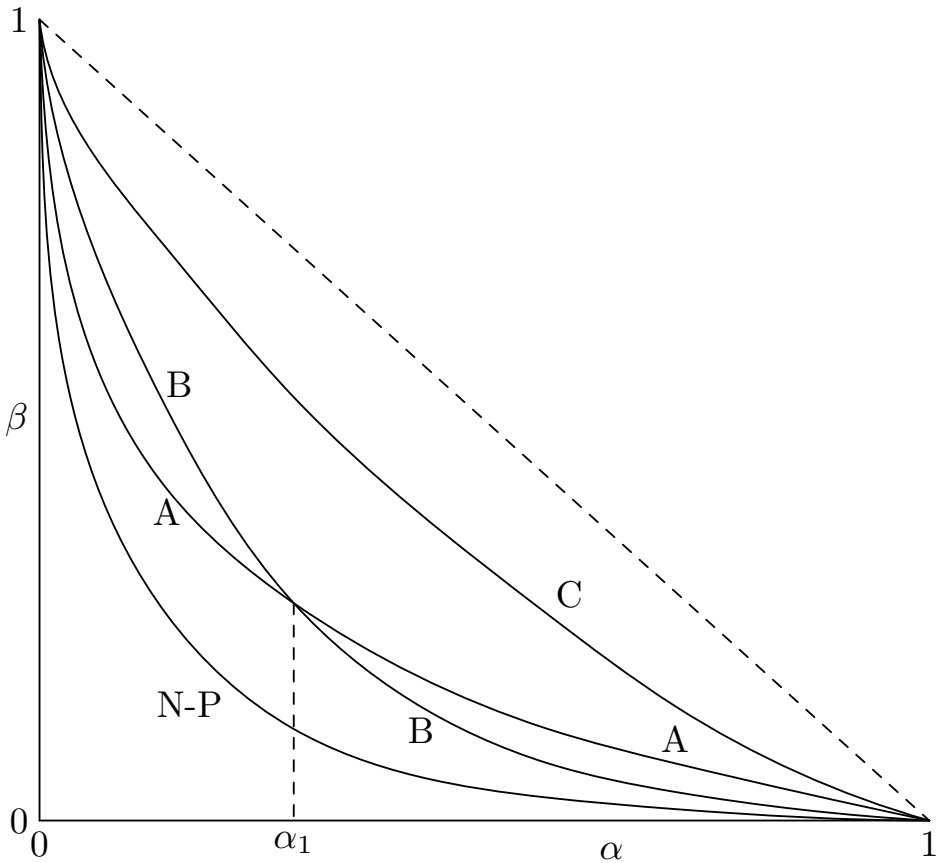
That is, we are more likely to accept the null hypothesis when it is false than when it is true. Such a test is called a *biased* test.



Power functions for biased (B) and unbiased (A) tests.

Choice of Tests

For a given test, one must still choose a value of α or β.
It is instructive to look at different tests in the $\alpha - \beta$ plane.



Unbiased tests must lie below the dotted line.
N-P is the most powerful test.

The Neyman-Pearson Test

Of all tests for H_0 against H_1 with significance α, the most powerful test is that with the best critical region in X-space, that is, the region with the smallest value of β.

Suppose that the random variable $\mathbf{X} = (X_1, \ldots, X_N)$ has p.d.f. $f_N(\mathbf{X}|\theta_0)$ under θ_0, and $f_N(\mathbf{X}|\theta_1)$ under θ_1.

From the definitions of α and the power $(1-\beta)$, we have

$$\int_{w_\alpha} f_N(\mathbf{X}|\theta_0) d\mathbf{X} = \alpha$$

$$1-\beta = \int_{w_\alpha} f_N(\mathbf{X}|\theta_1) d\mathbf{X}\,.$$

$$\begin{aligned} 1-\beta &= \int_{w_\alpha} \frac{f_N(\mathbf{X}|\theta_1)}{f_N(\mathbf{X}|\theta_0)} f_N(\mathbf{X}|\theta_0) d\mathbf{X} \\ &= E_{w_\alpha}\left(\frac{f_N(\mathbf{X}|\theta_1)}{f_N(\mathbf{X}|\theta_0)} \,\middle|\, \theta = \theta_0 \right)\,. \end{aligned}$$

The Neyman-Pearson Test

$$1 - \beta = E_{w_\alpha}\left(\frac{f_N(\mathbf{X}|\theta_1)}{f_N(\mathbf{X}|\theta_0)} \bigg|_{\theta=\theta_0} \right)$$

Clearly this will be maximal if and only if w_α is that fraction α of X-space containing the largest values of $f_N(\mathbf{X}|\theta_1)/f_N(\mathbf{X}|\theta_0)$. Thus the best critical region w_α consists of points satisfying

$$\ell_N(\mathbf{X}, \theta_0, \theta_1) \equiv \frac{f_N(\mathbf{X}|\theta_1)}{f_N(\mathbf{X}|\theta_0)} \geq c_\alpha \,,$$

The procedure leads to the criteria:

$$\begin{array}{llll} \text{if} & \ell_N(\mathbf{X}, \theta_0, \theta_1) > c_\alpha & \text{choose} & H_1 : \; f_N(\mathbf{X}|\theta_1) \\ \text{if} & \ell_N(\mathbf{X}, \theta_0, \theta_1) \leq c_\alpha & \text{choose} & H_0 : \; f_N(\mathbf{X}|\theta_0) \,. \end{array}$$

This is the Neyman–Pearson test. The test statistic ℓ_N is essentially the ratio of the likelihoods for the two hypotheses, and this ratio must be calculable at all points $\mathbf{X}$ of the observable space. The two hypotheses H_0 and H_1 must therefore be completely specified simple hypotheses, and then this gives the best test.

Composite Hypotheses

The theory above applies only to simple hypotheses.

Unfortunately, we often need to test hypotheses which contain unknown free parameters, composite hypotheses, such as:

$$\begin{aligned} H_0: &\quad \theta_1 = a, &\quad \theta_2 = b \\ H_1: &\quad \theta_1 \neq a, &\quad \theta_2 \neq b \end{aligned}$$

or

$$\begin{aligned} H_0: &\quad \theta_1 = a, &\quad \theta_2 \text{ unspecified} \\ H_1: &\quad \theta_1 = b, &\quad \theta_2 \text{ unspecified} \end{aligned}$$

Existence of Optimal Tests

For composite hypotheses there is in general no UMP test.

However, when the pdf is of the exponential form, there is a result similar to Darmois' Theorem for the existence of sufficient statistics.

If $X_1 \cdots X_N$ are independent, identically distributed random variables with a p.d.f. of the form

$$F(X)G(\theta)\exp[A(X)B(\theta)]\,,$$

where $B(\theta)$ is strictly monotonic, then there exists a UMP test of

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta > \theta_0.$$

(Note that this test is only one-sided)

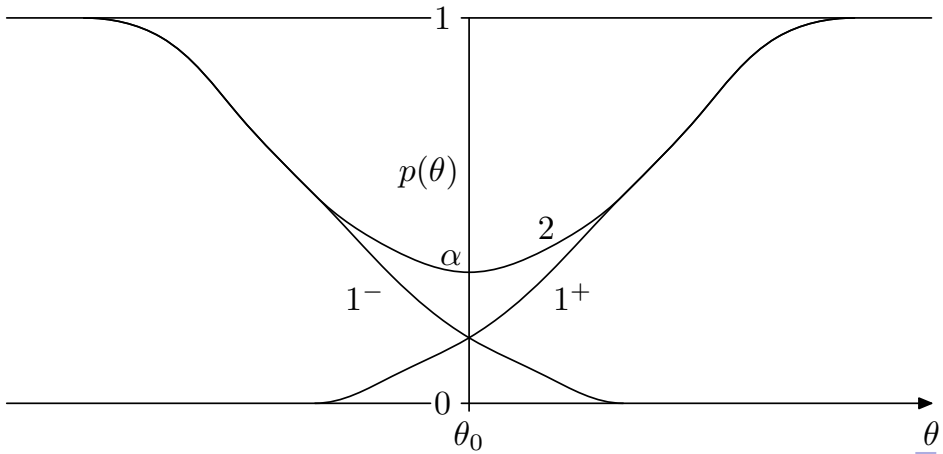
One-sided and two-sided tests

When the test involves the value of one parameter θ, we can have a one-sided test of the form $H_0 : \theta = \theta_0$ $H_1 : \theta > \theta_0$
or a two-sided test of the form $H_0 : \theta = \theta_0$ $H_1 : \theta \neq \theta_0$

For two-sided tests, no UMP test generally exists as can be seen from the power curves shown below:

- Test 1^+ is UMP for $\theta > \theta_0$
- Test 1^- is UMP for $\theta < \theta_0$
- Test 2 is the (two-sided) sum of tests 1^+ and 1^-.

At the same significance level α, Test 2 is clearly less powerful than Test 1^+ on one side and Test 1^- on the other.



Maximizing Local Power

If no UMP test exists, an important alternative is to look for a test which is most powerful in the neighbourhood of the null hypothesis. Then we have

$$H_0 : \theta = \theta_0 \quad H_1 : \theta = \theta_0 + \Delta, \qquad (\Delta \text{ small})$$

Expanding the log-likelihood we have

$$\ln L(\mathbf{X}, \theta_1) = \ln L(\mathbf{X}, \theta_0) + \Delta \left.\frac{\partial \ln L}{\partial \theta}\right|_{\theta=\theta_0} + \cdots .$$

If we now apply the Neyman–Pearson lemma to H_0 and H_1, the test is of the form:

$$\ln L(\mathbf{X}, \theta_1) - \ln L(\mathbf{X}, \theta_0) \gtrless c_\alpha .$$

or

$$\left.\frac{\partial \ln L}{\partial \theta}\right|_{\theta=\theta_0} \gtrless k_\alpha .$$

maximizing local power (cont.)

If the observations are independent and identically distributed, then

$$E\left(\left.\frac{\partial \ln L}{\partial \theta}\right|_{\theta=\theta_0}\right) = 0$$

$$E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = +NI\,,$$

where N is the number of observations and I is the information matrix. Under suitable conditions $\partial \ln L/\partial\theta$ is approximately Normal. Hence a locally most powerful test is approximately given by

$$\left.\frac{\partial \ln L}{\partial \theta}\right|_{\theta=\theta_0} \gtrless \lambda_\alpha \sqrt{NI}\,.$$

Likelihood Ratio Test

This is the extension of the Neyman-Pearson Test to composite hypotheses. Unfortunately, its properties are known only asymptotically.

Let the observations $\mathbf{X}$ have a distribution $f(\mathbf{X}|\boldsymbol{\theta})$, depending on parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$. Then the likelihood function is

$$L(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(X_i|\boldsymbol{\theta})\,.$$

In general, let the total $\boldsymbol{\theta}$-space be denoted θ, and let ν be some subspace of θ, then any test of parametric hypotheses (of the same family) can be stated as

$$H_0 : \boldsymbol{\theta} \in \nu$$

$$H_1 : \boldsymbol{\theta} \in \theta - \nu$$

Likelihood Ratio Test (cont.)

We can then define the maximum likelihood ratio, a test statistic for H_0:

$$\lambda = \frac{\max\limits_{\boldsymbol{\theta}\in\nu} L(\mathbf{X}|\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta}\in\theta} L(\mathbf{X}|\boldsymbol{\theta})} .$$

If H_0 and H_1 were simple hypotheses, λ would reduce to the Neyman-Pearson test statistic, giving the UMP test. For composite hypotheses, we can say only that λ is always a function of the sufficient statistic for the problem, and produces workable tests with good properties, at least for large sets of observations.

Likelihood Ratio Test (cont.)

The maximum likelihood ratio is then the ratio between
the value of $L(\mathbf{X}|\boldsymbol{\theta})$, maximized with respect to $\theta_j, j = 1, \ldots, s$,
while holding fixed $\theta_i = \theta_{io}$, $i = 1, \ldots, r$,
and the value of $L(\mathbf{X}|\boldsymbol{\theta})$, maximized with respect to *all* the parameters.
With this notation the test statistic becomes

$$\lambda = \frac{\max\limits_{\boldsymbol{\theta}_s} \; L(\mathbf{X}|\boldsymbol{\theta}_{r0}, \boldsymbol{\theta}_s)}{\max\limits'_{\boldsymbol{\theta}_r, \boldsymbol{\theta}_s} \; L(\mathbf{X}|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)}$$

$$\text{or} \qquad \lambda = \frac{L(\mathbf{X}|\boldsymbol{\theta}_{r0}, \boldsymbol{\theta}''_s)}{L(\mathbf{X}|\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_s)} \,. \qquad \longleftarrow \begin{pmatrix} \text{correct your book} \\ \text{p. 271, Eq(10.11)} \end{pmatrix}$$

where $\boldsymbol{\theta}''_s$ is the value of $\boldsymbol{\theta}_s$ at the maximum in the restricted $\boldsymbol{\theta}$ region and $\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_s$ are the values of $\boldsymbol{\theta}_r, \boldsymbol{\theta}_s$ at the maximum in the full $\boldsymbol{\theta}$ region.

Likelihood Ratio Test (cont.)

The importance of the maximum likelihood ratio comes from the fact that asymptotically:

if H_0 imposes r constraints on the $s + r$ parameters in H_0 and H_1, then

$$-2 \ln \lambda \text{ is distributed as } \chi^2(r) \text{ under } H_0$$

This means we can read off the confidence level α from a table of χ^2.

However, the bad news is that this is only true asymptotically,
and there is no good way to know how good the approximation is
except to do a Monte Carlo calculation.

Likelihood Ratio Test - Example

Problem: Find the ratio X of two complex decay amplitudes:

$$X = \frac{A(\text{reaction 1})}{A(\text{reaction 2})}.$$

In the general case, X may be any complex number, but there exist three different theories which predict the following for X:

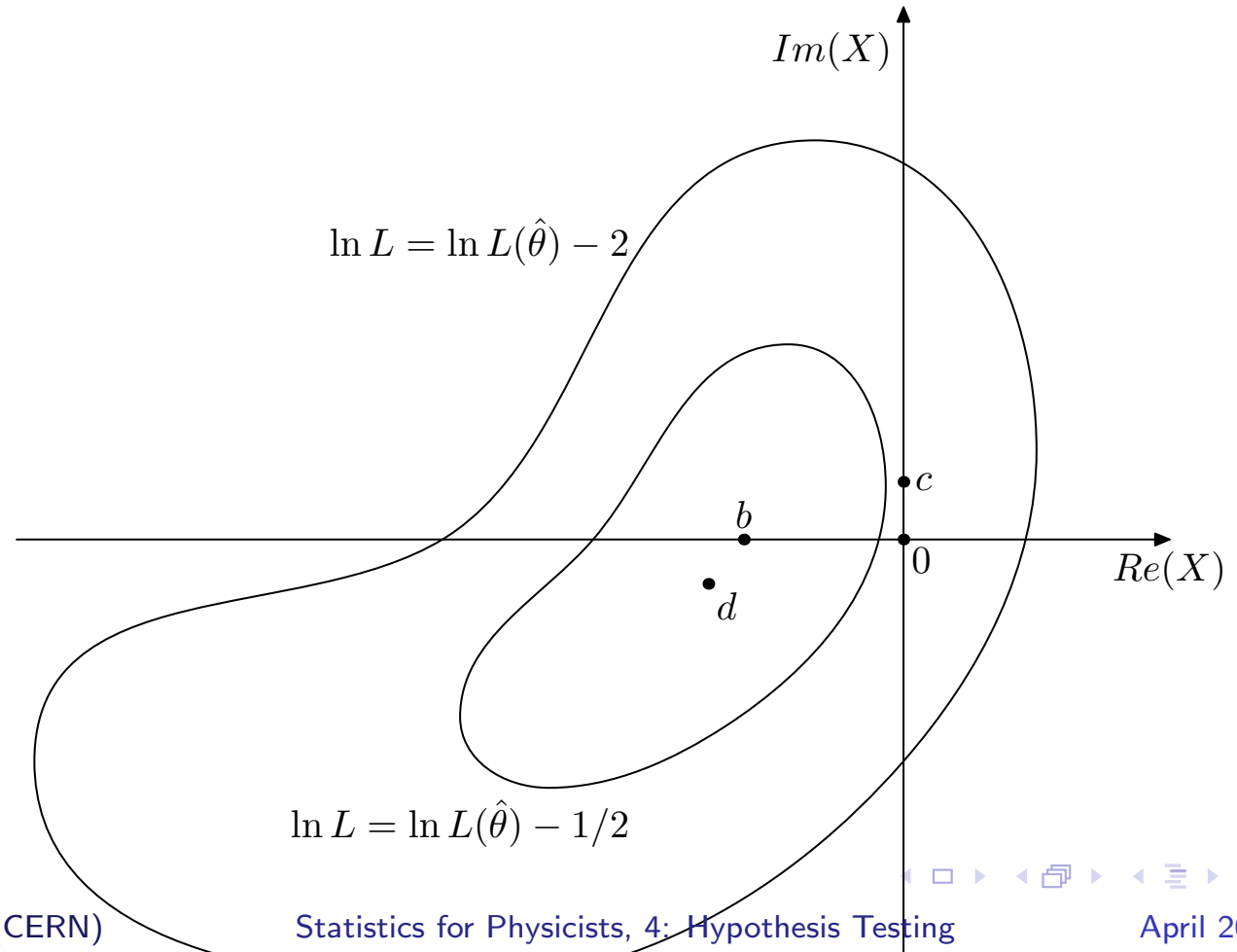
- A: If Theory A is valid, $X = 0$.
- B: If Theory B is valid, X is real and $Im(X) = 0$.
- C: If Theory C is valid, X is purely imaginary and non-zero.

We decide that the value of X is interesting only in so far as it could distinguish between the hypotheses A, B, C or the general case.
Therefore, we are doing hypothesis testing, not parameter estimation.

Hypothesis A is simple,
Hypothesis B is composite, including hypothesis A as a special case.
Hypothesis C is also composite, and separate from A and B.
The alternative to all these is that Re (X) and Im (X) are both non-zero.

Likelihood Ratio Test - Example

The contours of the log-likelihood function $\ln L(X)$ near its maximum.
$X = d$ is the point where $\ln L$ is maximal.
$X = b$ is the maximum of $\ln L$ when $Im(X) = 0$.
$X = c$ is the maximum of $\ln L$ when $Re(X) = 0$.



Likelihood Ratio Test - Example

The maximum likelihood ratio for hypothesis A versus the general case is

$$\lambda_a = \frac{L(0)}{L(d)}\,.$$

If hypothesis A is true, $-2\ln\lambda_a$ $\longleftarrow$ correct book p.276, line 5 is distributed asymptotically as a $\chi^2(2)$, and this give the usual test for Theory A.
To test Theory B, the m.l. ratio for hypothesis B versus the general case is

$$\lambda_b = \frac{L(b)}{L(d)}\,.$$

If B is true, $-2\ln\lambda_b$ is distributed asymptotically as a $\chi^2(1)$. Finally, Theory C can be tested in the same way, using $L(c)$ in place of $L(b)$.

Hypothesis Testing - Bayesian

Recall that according to Bayes' Theorem:

$$P(\text{hyp}\,|\,\text{data}) = \frac{P(\text{data}\,|\,\text{hyp})P(\text{hyp})}{P(\text{data})}$$

The normalization factor $P(\text{data})$ can be determined for the case of parameter estimation, where all the possible values of the parameter are known, but in hypothesis testing it doesn't work, since we cannot enumerate all possible hypotheses. However it can be used to find the **ratio of probabilities** for two hypotheses, since the normalizations cancel:

$$R = \frac{P(H_0\,|\,\text{data})}{P(H_1\,|\,\text{data})} = \frac{\mathcal{L}(H_0)P(H_0)}{\mathcal{L}(H_1)P(H_1)}$$

Hypothesis Testing - Bayesian

But how do we interpret the ratio of two probabilities?
This would be obvious for frequentist probabilities, but what is a ratio of beliefs?

If $R = 2.5$, for example, it means that we believe in H_0 2.5 times as much as we believe in H_1.
How can we use the value of R?

For betting on H_0. It gives us directly the odds we can accept if we want to bet on H_0 against H_1.

Note that R is proportional to the ratio of prior probabilities $P(H_0)/P(H_1)$, so there is no way to make the result insensitive to the priors.